

end
a1/

BACKGROUND OF THE INVENTION

1. Field of the Invention

5

This invention is related to the field of computer networks and, more particularly, to maintaining the high availability in a two node computer network which utilizes alternate pathing technology.

10 2. Description of the Related Art

With the ever expanding use of computer networks throughout society has come an increasing dependence of users on the availability of that network. If a network goes down, or is otherwise unavailable, costs to an enterprise may be significant.

15 Consequently, a number of techniques have arisen which are designed to ensure that a computer network is sufficiently robust that it may detect and respond to problems without significantly impacting users. Frequently, efforts to ensure a computer network is consistently online for its users may be referred to as maintaining "high availability". A computer network which has in place mechanisms which prevent hardware or software
20 problems from impacting its users may be referred to as a High Availability Network (HANet). Some of the characteristics which may be considered when defining a HANet include protection of data (Reliability), continuous access to data (Availability), and techniques for correcting problems which minimally impact users (Serviceability). Collectively these characteristics are frequently referred to as RAS.

25

In some cases it is desirable to create a computer network which includes a two node Local Area Network (LAN). For example, it may be desirable to have a two node LAN consisting of a database server and its corresponding application server. These servers may be connected to each other using a well known method using crossover

cables. A crossover cable is a cable that is used to connect two computers by reversing their respective pin contacts. Using crossover cables may have the advantage of being highly secure, performing well, and eliminating several components typically present in a computer network, such as switches and routers, which could cause a failure. However, while such a configuration may improve reliability and availability in the system, it does not address serviceability and still contains single points of failure. For example, failure of either of the server network interfaces to which the crossover cable is connected will cause the network to be unavailable. Also, failure of the crossover cable itself would result in unavailability of the network. In some cases, mechanisms may be put in place which detect an error in a network connection and notify the system administrator that a problem exists. The system administrator may then take corrective action, such as switching to a redundant resource. However, such mechanisms typically take some period of time and necessarily involve interruptions in network operation. In other cases, operating system specific mechanisms may be implemented which may facilitate a failover to a redundant connection. Typically these mechanisms operate at layers below the application layer of the protocol stack. Two widely recognized protocols include TCP/IP and ISO/OSI, each of which include a highest layer referred to as the application layer. Other communication protocols with a layer corresponding to the application layer may utilize a different name. Generally, those layers below the application layer involve software and mechanisms which are not portable across different operating systems. Consequently, these solutions are not portable and generally require a newly created mechanism for each platform on which a failover is desired.

One technology which provides for redundancy in case of failure is alternate pathing. Alternate pathing is a technology which provides for redundancy to storage in case of a failed I/O controller. In addition to providing for recovery after failure, alternate pathing may also be used to support dynamic reconfiguration. Dynamic reconfiguration is used to logically attach and detach system boards from a running operating system. In addition to providing redundancy to storage, alternate pathing may also be used with

network connections. However, alternate pathing does not support automatic failover for network connections. Consequently, the problems described above still remain.

SUMMARY OF THE INVENTION

5

The problems outlined above are in large part solved by a method and mechanism as described herein. A method and mechanism of failover in a system with alternate pathing is described. By utilizing an Application layer mechanism which monitors the primary network connection, automatically detects a failure in the primary connection, and switches to the secondary connection in a short period of time, network availability may be maintained. Advantageously, network interruptions may be minimized and servicing of network problems may be automated by a mechanism which is portable across multiple platforms. Further, because the mechanism operates within the application layer of the communication protocol, no modification of existing operating software is necessary.

Broadly speaking, a method for maintaining high availability in a two node computer network utilizing alternate pathing is contemplated. The method includes adding an Application layer High Availability Networking (HANet) mechanism to a node of the computer network, monitoring a first network connection, detecting a failure of the first network connection, and performing a failover from the first network connection to the second network connection. The monitoring, failure detection, and failover are all performed by the HANet mechanism.

Also contemplated is a network node configured to support alternate pathing which includes a first network interface, a second network interface, and a High Availability Networking (HANet) mechanism. The included HANet mechanism operates at the Application layer and is configured to monitor the first network interface. If a

failure of the first network interface is detected, the HAnet mechanism is configured to perform a failover from the first network interface to the second network interface.

Further contemplated is a two node computer network configured to support alternate pathing and to maintain high availability. The network includes a first node coupled to a second node by two paths. The first node includes a High Availability Networking (HAnet) mechanism which operates at the Application layer. The HAnet mechanism is configured to monitor the first path and perform a failover from the first path to the alternate path in response to detecting a failure of the first path.

BRIEF DESCRIPTION OF THE DRAWINGS

Other objects and advantages of the invention will become apparent upon reading the following detailed description and upon reference to the accompanying drawings in which:

Fig. 1 is a diagram of a computer network.

Fig. 2 is a diagram of a two node local area network.

Fig. 3 is a diagram of a two node network with a redundant connection.

Fig. 4 is a diagram of a two node network with a failover mechanism.

Fig. 5 is a diagram of a flowchart illustrating a failover method.

Fig. 6 is a diagram of a computer network with a firewall.

Fig. 7 is a diagram of a computer network with a firewall and failover mechanism.

Fig. 8 is a diagram of a computer network with trunking.

Fig. 9 is a diagram of a computer network with trunking and a failover
5 mechanism.

Fig. 10 is a diagram of a flowchart illustrating a failover method.

Fig. 11 is a diagram of a computer network with alternate pathing.
10

Fig. 12 is a diagram of a computer network with alternate pathing and a failover
mechanism.

Fig. 13 is a diagram of a computer network utilizing virtual IP addresses and a
15 failover mechanism.

While the invention is susceptible to various modifications and alternative forms,
specific embodiments thereof are shown by way of example in the drawings and will
herein be described in detail. It should be understood, however, that the drawings and
20 detailed description thereto are not intended to limit the invention to the particular form
disclosed, but on the contrary, the intention is to cover all modifications, equivalents and
alternatives falling within the spirit and scope of the present invention as defined by the
appended claims.

25 **DETAILED DESCRIPTION OF THE INVENTION**

Computer Network Overview

Fig. 1 is an illustration of a computer network 100. Included in computer network 100 are database server 104, application server 102A, application server 102B, network infrastructure 108, and workstation 110. Database server 104 is coupled to application servers 102A and 102B. Application server 102A and 102B are further coupled to network infrastructure 108. Workstation 108 is also coupled to network infrastructure 108. In addition, database server 104 includes network interfaces 120B and 120C, application server 102A includes network interfaces 120A and 120F, and application server 102B includes network interfaces 120D and 120E. Elements referred to herein with a particular reference number followed by a letter will be collectively referred to by the reference number alone. For example, interfaces 120A-120F will be collectively referred to as interfaces 120. In one embodiment, interfaces 120 are network interface cards (NIC).

In computer network 100 of ^{Fig} ~~fig.~~ 1, users may access the network 100 via workstations, such as workstation 110, coupled to network infrastructure 108. In one embodiment, a user on a workstation 110 may interact directly with a "front-end" interface, which in turn communicates with an application on application server 102A. The application on application server 102A may then access required data from data base server 104. Application server 102A may be viewed as in intermediary between the front-end workstation 110 and the back-end database 104. By acting as an intermediary, application server 102A prevents users from directly accessing database server 104 and may better balance the load on database server 104. While there are advantages to utilizing the two node structure of application server 102A and database server 104, having a single connection between the two includes the existence of "single points of failure" as discussed in ^{Fig} ~~fig.~~ 2. A single point of failure may be described as a point in the network, for example a device or connection, which has no backup and whose failure will cause an interruption in network service.

Fig. 2 is an illustration of a two node network. Included in ^{Fig}fig. 2 are database server 202 and application server 204. Database server 202 includes NIC 220B. Application server 204 includes NIC 220A and 220C. In addition, application server 204 includes Chinese font application 260. NIC 220A is coupled to NIC 220B via interconnect 240. In one embodiment, interconnect 240 is a crossover cable. In the example of ^{Fig}fig. 2, a user may access application server 204 via interconnect 250. A user may require access to application server 204 for the purpose of utilizing Chinese font application 260. In communicating with a user, application server 204 may access database server 202.

As shown in ^{Fig}fig. 2, there exists a single interconnect 240 between application server 204 and database server 202. If that single interconnect 240 fails, communication between application server 204 and database server 202 will be lost. Likewise, if either interface NIC 220A or NIC 220B fails, communication between application server 204 and database server 202 will be lost. Consequently, each of these components 240, 220A and 220B represents a single point of failure as described above. Further, because there is no backup connection between server 202 and 204, if this single connection 230 fails the only solution is to replace or repair the defective component. In any event, the failure of this single connection 230 will result in interruption of network service. In order to address this problem, a redundant connection may be introduced as discussed in the following.

Fig. 3 is an illustration of a two node network which includes a redundant connection. Included in the network are a primary connection 330A and a secondary connection 330B. Primary connection 330A includes NIC 320A, NIC 320B, and interconnect 340. Secondary connection 330B includes NIC 322A, NIC 322B, and interconnect 342. In one embodiment, primary connection 330A serves to transmit communications between database server 202 and application server 204. If a failure of connection 330A occurs, connection 330B may be utilized to communicate between

servers 202 and 204. By incorporating secondary connection 330B, the single point of failure that exists in a single connection network as described above is eliminated.

However, in order to avoid interruptions in network service, secondary connection 330B must be able to assume the functions of primary connection 330A in a very short period of time. If a failure of primary connection 330A occurs and secondary connection 330B does not assume the functions of primary connection 330A very rapidly, users will experience network service interruption. In the following discussion, a failover method is described which provides for high availability and serviceability.

10 Network Protocol and Addressing Overview

A well known communications protocol for networking is the Transmission Control Protocol/Internet Protocol (TCP/IP) communications protocol. In general, each host connected to a network has an assigned logical address, or Internet Protocol (IP) address. When one host wishes to convey data to another host, packets are formed which contain the data and the IP address of the destination host and the packets are then delivered to the desired destination. Necessarily, each host must have a unique IP address. Each IP address is a sequence of digits comprised of two basic parts, a network ID and a host ID. The network ID identifies devices which are located on the same physical network. The host ID is used to identify a particular host within a network. A second address which may be associated with a host is a subnet mask. A subnet mask is used to divide a larger network into smaller sub-networks. Frequently, network administration may be simplified by carefully defining sub-networks for a network. A subnet mask includes a sequence of digits similar to an IP address and is used to indicate which part of an IP address corresponds to a network ID and which part corresponds to host ID. A third address used in the IP protocol is a broadcast address. The broadcast address may be used by a host to convey data to all other hosts on a network simultaneously, as opposed to sending data to a targeted host with its unique IP address. Finally, each host interface includes a physical, or Media Access Control (MAC), address. The MAC address is

typically configured at the time of manufacture of the interface. In one embodiment, interface MAC addresses may be specified by the host of which they are a part. In such a case, a host with multiple interfaces may specify the same MAC address for all interfaces. In an alternative embodiment, each interface on a given host may have a different MAC address. Because packet delivery is handled at the physical layer and IP addresses are only logical addresses, IP addresses must be converted to corresponding MAC addresses. IP uses the well known Address Resolution Protocol (ARP) to build a table in memory which maps IP addresses to corresponding MAC addresses. By referencing this table, physical addresses corresponding to IP addresses may be determined.

10

In the ^{UNIX™} operating system, network interface configuration may be done using the well known “ifconfig” command. Among other things, the ifconfig command allows the configuration of an interface’s IP address, subnet mask, and broadcast address. The basic syntax for ifconfig is typically as follows:

15

ifconfig [interface] [command] [address] [options]

20

where “interface” is the name of the interface, “command” includes a variety of command which depends on the particular version of ^{UNIX™} being run, “address” is either a host name or address expressed in typical dot notation, and “options” may be any of a number of configuration options. Some of the options which may be specified are “up”, “down”, “netmask”, “broadcast”, and “plumb”. The plumb option is used to setup and open the streams necessary for TCP/IP to use the interface. The broadcast option is used to configure a broadcast address. The netmask option is used to establish the subnet mask for the interface. The down option prevents the system from communicating through the interface. Finally, the up option enables the interface for communication. For example, the following command may be used to assign the IP address 172.16.1.2 to the interface named le0:

30

ifconfig le0 inet 172.16.1.2.

a

UNIX 74

Another well known utility in ~~UNIX~~ is the “ping” command. The ping command uses timed IP/ICMP ECHO_REQUEST and ECHO_REPLY packets to probe the

5 "distance" to a target machine. By examining packets received from the target machine, communication latency may be determined. In one embodiment, the ping utility may be used to determine if a particular network connection is operational. This use is sometimes referred to as a “heartbeat”, for it indicates whether or not a particular node or connection is “alive”. For example, one node on a network may use the ping command to

10 periodically send requests to another node over a particular network interface. As long as a corresponding reply is received, it may be assumed that the connection is working. If no reply is received, the connection may no longer be operational and corrective action may be required. In addition to the ping utility, other similar types of request/reply communication utilities may be used to test network connections as well. As an

15 alternative to using the ping utility, some NICs are configured to maintain a status of the network connection to which they are connected. Such a status may be indicated by a status register within the NIC itself. By polling this status register, such as through device driver calls, a node may detect whether or not a particular connection is operational. By utilizing a network interface configuration utility such as “ifconfig”, and monitoring a

20 network connection, a mechanism for maintaining high availability may be created as described by the following.

Fig. 4 is a diagram of one embodiment of a computer network including host 402, host 404, primary connection 450 and secondary connection 452. Host 402 includes

25 interfaces hme0 and hme1, NIC 420, NIC 422, and failover mechanism 410A. In one embodiment, failover mechanism 410 includes software running at the application layer of the communications protocol. Host 404 includes NIC 440, NIC 442 and failover mechanism 410B. Interface hme0 is a primary interface and hme1 is a secondary interface. NIC 420 is coupled to NIC 440 via connection 430 and NIC 422 is coupled to

corresponding to NIC 420. Likewise, failover mechanism 410B may detect a failure in primary connection 450 and perform a failover to secondary connection 452. Failover mechanism 410A may include the following commands to perform the failover:

5 **pause_monitor**

ifconfig hme0 unplumb

ifconfig hme1 plumb

ifconfig hme1 inet 129.1.1.1 netmask 255.255.255.0 broadcast 129.1.1.255 up

resume_monitor

10

Upon detecting a failure, the failover routine is called and monitoring of the primary connection 450 is paused. The first ifconfig command above “unplumbs” interface hme0 so that host 402 will not attempt to use it for communications. The second ifconfig command above “plumbs” interface hme1 and sets up the TCP/IP streams needed for host 402 to use this interface. The third ifconfig command above configures interface hme1 with the parameters corresponding to hme0. Interface hme1 is configured with the IP address 129.1.1.1, subnet mask 255.255.255.0, and the broadcast address 129.1.1.255. The “up” parameter brings interface hme1 up so that it may be used by host 402 for communications. Upon completion of the failover to the new interface, monitoring of the network connection is resumed. Because NIC 420 and NIC 422 have the same MAC address, there is no need to transfer the MAC address of NIC 420 to NIC 422. Failover mechanism 410B may utilize a similar set of commands to perform failover to secondary connection 452. Advantageously, the connection failure is transparent to clients of the node and the network connection is maintained. Further, because failover mechanism 410 operates at the application layer, it is not necessary to modify the operating system, drivers, or network software of host 402 or host 404. Advantageously, failover mechanism 410 is portable and may be easily implemented in a wide variety of systems running a variety of operating systems. In one embodiment, upon completing a failover to secondary connection 452, monitoring of secondary connection 452 begins and a failover

to primary connection 450 may be initiated in the event secondary connection 452 fails. Further, in one embodiment it may be necessary to “unplumb” primary connection 450 prior to plumbing secondary connection 452.

5 In one embodiment, failover mechanism 410 may include a configuration file specifying a delay parameter. The delay parameter may indicate a specified period of time a host is to wait before checking network connectivity. In this manner, both hosts 402 and 404 may have time to perform failover before one begins communicating again. Were one host to begin communicating prior to completion of failover by both, the
10 network connection may appear inoperable to the transmitting host and another failover may be initiated.

Fig. 5 is a flowchart illustrating a failover method. Initially, a secondary network connection is added to a two node LAN (block 502). This newly added connection
15 includes an additional NIC added to each node and a crossover cable coupling the new NICs together. The primary network connection is then monitored for failure (block 504). Such monitoring may include using a ping utility or polling a status register of the corresponding NIC. If no failure of the primary connection is detected (decision block 506), monitoring of the primary connection continues (block 504). If a failure of the
20 primary connection is detected (decision block 506), monitoring of the primary connection is stopped (block 508), the primary connection is brought down (block 510) and the secondary connection is plumbed (block 512). Finally, the secondary connection is configured with the parameters of the primary connection and is brought up (block 514), and monitoring of the newly enabled connection is started (block 516).

25

Firewalls

A firewall is a well known technology which provides for restricted communications between separate networks. Firewalls are frequently used by an

enterprise to control access of those on an outside network, such as the Internet, to the enterprise's inner network. Firewalls may also be used to protect some parts of an inner network from other parts of an inner network. Firewalls typically use a complex set of predefined rules which are applied to communications between networks. Firewalls
5 generally attempt to match an incoming or outgoing communication and service with one of the predefined rules. If a rule match is found, the communication may be either accepted or rejected.

Fig. 6 is a diagram illustrating a computer network with a firewall. Fig. 6 includes
10 an external network 690, firewall 602, switch 604, and internal network 692. Firewall 602 includes NIC 620 and NIC 680. Switch 604 includes port 640 and port 682. External network 690 is coupled to firewall 602 through NIC 680. Internal network 692 is coupled to switch 604 through NIC 682. NIC 620 is coupled to port 640 via interconnect 630. In general, firewall 602 controls communications between the external network 690 and the
15 internal network 692. In a typical configuration, internal network 692 may be a LAN within a company and external network 690 may represent a network outside the company. By placing the firewall between the external network 690 and the internal network 692, communications may be controlled and security enhanced. As shown in ~~fig. 6~~^{Fig. 19}, there is a single connection 650 between firewall 602 and switch 604. If connection
20 650 were to fail, communication between internal network 692 and external network 690 would be lost. Connection 650 represents a single point of failure. To eliminate this single point of failure and provide for high availability, a failover mechanism as discussed below may be utilized.

25 Fig. 7 is a diagram illustrating a computer network with a firewall and an added redundant connection. Fig. 7 includes an external network 690, firewall 602, switch 604, switch 605 and internal network 692. Firewall 602 includes failover mechanism 610, NIC 620, NIC 622 and NIC 680. Switch 604 includes port 640 and port 682. Switch 605 includes ports 642 and 684. External network 690 is coupled to firewall 602 through NIC

680. Internal network 692 is coupled to switch 604 through port 682 and switch 605 through port 684. NIC 620 is coupled to port 640 via interconnect 630. NIC 622 is coupled to port 642 via interconnect 632. Switches 604 and 605 are coupled to each other. Connection 650 represents a primary connection between firewall 602 and switch 604. Connection 652 represents a secondary connection between firewall 602 and switch 604.

Because of the security oriented nature of firewalls, certain operations which may be permitted on a non-firewall server may not be permitted on a firewall server. For example, in certain firewall implementations, all interfaces must be defined when the system is started and all interfaces must be kept defined at all times in order to prevent security breaches. In such an implementation, plumbing an interface after boot may be seen as a hostile act by a firewall server and may cause the firewall server to shut down. Consequently, if the interface corresponding to NIC 622 is plumbed after firewall 602 has started, firewall 602 may shut down all communications between external network 690 and internal network 692. Further, if interfaces corresponding to both NIC 620 and 622 are plumbed at startup, they may both have the same MAC address which may not be permissible.

In order to implement a failover mechanism for a system such as firewall 602, the following method may be used. A redundant connection 652 is added as described above. On startup of firewall 602, failover mechanism 610 defines all interfaces. Failover mechanism 610 plumbs and configures NIC 620, NIC 622, and NIC 680. NIC 620 and NIC 680 are plumbed and configured with the actual parameters to be used in network communication. However, NIC 622 is plumbed and configured with dummy IP addresses and a dummy MAC address. Upon configuring NIC 622 with the dummy parameters, NIC 622 is immediately brought down by failover mechanism 610. In this manner, all interfaces may be defined at startup, but there is no conflict between IP addresses or MAC addresses. In this case, the secondary connection may be deemed a "warm" standby

as it is already plumbed and configured, but not "up" or active. As shown in ^{Fig}~~fig~~ 7, the following ^{exemplary} interface parameters are indicated:

Interface	NIC 620	NIC 622
Parameters	IP Address : 129.1.1.1 Netmask : 255.255.255.0 Broadcast : 129.1.1.255 MAC : 08-A0-20-CD-10-53	IP Address : 129.1.1.5 Netmask : 255.255.255.0 Broadcast : 129.1.1.255 MAC : 08-A0-20-CD-10-55

- 5 In order to perform the failover from primary connection 650 to secondary connection 652, a method similar to that discussed in ^{Fig}~~fig~~ 4 may be used. Utilizing a heartbeat or similar mechanism, a failure of the primary connection 650 may be detected by failover mechanism 610. Assuming failover mechanism 610 initially detects the failure of primary connection 650, failover mechanism 610 stops monitoring primary
- 10 connection 650. NIC 620 is then brought down. NIC 622 is then configured using ifconfig with the parameters which are currently assigned to NIC 620 and NIC 620 is configured with the dummy parameters which were associated with NIC 622. NIC 622 is then brought up. Failover mechanism 610 may then monitor connection 652 as the new primary connection and have connection 650 serve as the new secondary connection. In
- 15 the event connection 652 fails, a failover to connection 650 may be done. In alternate embodiments, firewall 602 may be coupled to another host which has a failover mechanism. In such a case, as discussed above, a configuration file may be utilized which includes a delay parameter to ensure both firewall 602 and the coupled host have completed failover before resuming communications. Advantageously, network
- 20 configurations which include security mechanisms such as firewalls may include the above described failover mechanism and high availability may be attained.

Increasing Bandwidth

Because of the ever increasing demands placed on networks today, ways of increasing network bandwidth are of frequent concern. While Fast Ethernet and Gigabit Ethernet may serve to improve performance, the use of such technologies necessitate the need for an even greater increase in backbone capacity. One well known technique used to increase bandwidth is called "trunking". Trunking is a technology which may provide dramatic increases in network performance. Using trunking technology, multiple ports may be combined into a single logical port creating a single, high speed, logical link. Fig. 8 is an illustration of a network utilizing trunking to boost throughput and reduce network latency. Included in fig. 8 are host 802, switch 820, LAN 1 830A and LAN 2 830B. Host 802 includes a four port NIC 880. NIC 880 is coupled to switch 820 via links 810A-810D. In one embodiment, each of links 810A-810D represent Fast Ethernet connections. Switch 820 is further coupled to LAN 1 830A and LAN 2 830B.

While NIC 880 has four ports, all four ports may be configured to act as a single logical port. Consequently, switch 820 may utilize any of the four ports to convey data to host 802. In one embodiment, connections 810 may be utilized in a round robin order. In this manner, links 810 may be viewed as a single connection with four times the capacity of any individual connection 810A-810D. Because all four ports of NIC 880 may be viewed as a single logical port, data communicated from LAN 1 830A or LAN 2 830B to host 802 has four possible connections upon which to be conveyed. In addition to increasing the bandwidth between Host 802 and ~~Switch~~ ^{Switch} 820, the availability of additional connections provides for redundancy as well. If one of connections 810 were to fail, the failed connection may be disabled and the remaining connections may continue service. In this manner, availability of the connection between Host 802 and LANs 830 may be maintained.

While trunking provides for increased bandwidth and redundancy, there still exists single points of failure. Because all four ports of NIC 880 must be connected to the same switch 820, a single point of failure is introduced. For example, if ~~Switch~~ ^{Switch} 820 fails,

communication between Host 802 and LANs 830 will be lost. Further, if NIC 880 were to fail, communication between Host 802 and LANs 830 would be lost. In order to eliminate single points of failure and maintain high availability, a failover method as described in the following may be introduced. Fig. 9 is a diagram of a network which includes trunking and a failover mechanism. Included in ^{Fig} ~~fig~~ 9 are host 802, switches 820, 920, 930A, 930B, and LANs 830A and 830B. Host 802 includes NIC 880 coupled to switch 820, and NIC 980 coupled to switch ^{Switch} ~~Switch~~ 920. Switch 820 and Switch 920 are further coupled to switches 930. LANs 830 are each coupled to both switch 930A and 930B. Host 802 also includes failover mechanism 990. NIC 880, connection 810, and ^{Switch} ~~Switch~~ 820 represent a primary connection 982. NIC 980, connection 910, and ^{Switch} ~~Switch~~ 920 represent a secondary connection 984.

In ordinary operation, primary connection 982 handles all network traffic between host 802 and LANs 830. In one embodiment, NIC 980 is initially unplumbed and unconfigured. NIC 880 may be configured as described above using the ifconfig command. Failover mechanism 990 may be configured to monitor primary connection 810 using a heartbeat, status register polling or other similar technique. In one embodiment, failover mechanism 990 is configured to monitor the operational status of each of the four ports 810A-810D of NIC 880. If failover mechanism 990 detects a failure of one of connections 810A-810D, a failover to secondary connection 984 may be ~~commenced. In the embodiment shown Primary~~ commenced. Primary NIC 880 tr0 is configured with an IP address of 129.1.1.1, a Netmask of 255.255.255.0 and a Broadcast address of 129.1.1.255. Secondary NIC 980 is unplumbed. Both NIC 880 and NIC 980 have a MAC address of 08-A0-20-CD-10-53. Assuming NIC 880 has the designation tr0 and NIC 980 has the designation tr1, the following commands may be used to perform the failover.

tr_config stop tr0

tr_config define tr1

tr_config tr1 129.1.1.1 netmask 255.255.255.0 broadcast 129.1.1.255 up

tr_config start_trunk_monitor

Upon detecting a failure, tr_config stop tr0 stops failover mechanism 990 from monitoring primary connection 982 and unplumbs tr0. The second tr_config command above “plumbs” interface tr1 and sets up the TCP/IP streams needed for host 802 to use this interface. The third tr_config command above configures interface tr1 with the parameters corresponding to tr0. Interface tr1 is configured with the IP address 129.1.1.1, subnet mask 255.255.255.0, and the broadcast address 129.1.1.255. The “up” parameter brings interface tr1 up so that it may be used by host 802 for communications. Finally, the failover switch procedure returns and failover mechanism 990 begins monitoring secondary connection 984. Because NIC 880 and NIC 980 have the same MAC address, there is no need to transfer the MAC address of NIC 880 to NIC 980. Advantageously, the connection failure is transparent to clients of the node and the network connection is maintained. Further, because failover mechanism 990 operates at the application layer, it is not necessary to modify the operating system, drivers, or network software of host 802. Advantageously, failover mechanism 990 is portable and may be easily implemented in a wide variety of systems running a variety of operating systems. In one embodiment, upon completing a failover to secondary connection 984, monitoring of secondary connection 984 begins and a failover to primary connection 982 may be initiated in the event secondary connection 984 fails.

In an alternative embodiment, failover mechanism 990 may be configured to failover only when the performance level of a connection falls below a predetermined threshold. For example, while monitoring primary connection 982, failover mechanism 990 may detect a failure of connection 810A only. Trunking capability may then disable connection 810A and utilize the remaining three connections, 810B-810D, to continue communications. If the level of performance of primary connection 982 still exceeds a particular threshold, failover mechanism 990 may not perform a failover. If then failover mechanism 990 detects a failure of a second connection, 810B, and the performance of

the primary connection 982 falls below the threshold, a failover to secondary connection 984 may be performed. Advantageously, failover may be performed only when deemed necessary. In one embodiment, failover mechanism 990 may be configured to read a configuration file to determine the failover threshold. For example, a threshold parameter may be set to 50% of connection capacity. Failover mechanism 990 may then perform a failover if 50% or more of the connections in a trunked interface fail.

Fig. 10 is a flowchart illustrating a method for failover in a network utilizing trunking. Initially, a secondary network connection is added to a host (block 1002). This newly added connection includes an additional multi-ported NIC coupled to a trunking enabled switch. The primary network connection is then monitored for failure (block 1004). Such monitoring may include using a ping utility or polling a status register of the corresponding NIC. If no failure of the primary connection is detected (decision block 1006), monitoring of the primary connection continues (block 1004). If a failure of one or more of the trunking connections in the primary connection is detected (decision block 1006), a determination is made as to whether a performance related threshold parameter has been exceeded (decision block 1007). Such a determination may include reading a threshold parameter from a configuration file as described above. If the threshold parameter has not been exceeded, flow returns to block 1004. If the threshold parameter has been exceeded (decision block 1007), monitoring of the primary connection is paused (block 1008), the primary connection is brought down (block 1010) and the secondary connection is plumbed (block 1012). Finally, the secondary connection is configured with the parameters of the primary connection and brought up (block 1014), and monitoring of the newly enabled connection is started (block 1016).

In addition to the above described four port (quad) trunking technology, the described failover mechanism may be applied to other trunking technologies as well. For example, two Gigabit Ethernet ports may be combined to form a trunk. In the event one of the two Gigabit Ethernet connections fails, a failover may be performed to a redundant

Gigabit Ethernet trunk. Further, the above description assumes a secondary connection is initially unplumbed. However, in the case of a host which utilizes a firewall technology, or other technology requiring all interfaces be defined at startup, a warm standby as described above may be utilized.

5

Alternate Pathing

Alternate pathing is a technology which provides for redundancy to storage in case of a failed I/O controller. In addition to providing for recovery after failure, alternate pathing may also be used to support dynamic reconfiguration. Dynamic reconfiguration is used to logically attach and detach system boards from a running operating system, without the need to reboot. Fig. 11 is an illustration of a system utilizing alternate pathing technology. Included in ~~fig.~~^{Fig.} 11 are host 1102, I/O controllers 1110A and 1110B, and disk array 1140. I/O controller 1110A is coupled to disk array 1140 via interconnect 1120A. I/O controller 1110B is coupled to disk array 1140 via interconnect 1120B. During ordinary operation, I/O controller 1110A is used for all I/O transactions between host 1102 and disk array 1140. In the event I/O controller 1110A fails, I/O controller 1110B is utilized for transactions between host 1102 and disk array 1140. On the other hand, the dynamic reconfiguration feature of alternate pathing provides that if the system administrator desires to remove or replace I/O controller 1110A without stopping and restarting host 1102, I/O controller 1110B may be enabled and I/O controller 1110A removed without the need to reboot host 1102.

Alternate pathing may also be used with network interfaces. When used with network interfaces, alternate pathing utilizes "metanetworking" to support alternate paths. A metanetwork is a logical construct that enables you to access a network by using one of two physical paths without having to explicitly reference either path. You reference a metanetwork using a metanetwork interface name such as "mether1" as will be discussed below. However, while alternate pathing may provide for dynamic reconfiguration of

network interface cards, automatic failover for networks after boot is not supported. If alternate pathing is used with network interfaces and a failure occurs, a manual failover is required. Because of the necessary delay involved in a manual failover, network operations will be interrupted. By adding a failover mechanism to a system with alternate pathing, automatic failover and high availability may be attained.

Fig. 12 is a diagram illustrating a system with alternate pathing and a failover mechanism. Shown in fig. 12 is a host 1202 coupled to an Ethernet interconnect 1270. Host 1202 includes NIC 1220, NIC 1222 and failover mechanism 1210. NIC 1220 is coupled to Ethernet 1270 via connection 1230 and NIC 1222 is coupled to Ethernet 1270 via connection 1232. Connection 1250 represents a primary path and connection 1252 represents an alternate path. While NIC 1220 is named hme0 and NIC 1222 is named hme1, alternate pathing provides that host 1202 may refer to either path using the metaname "mether1". Primary path 1250 is active and alternate path 1252 is inactive. Consequently, all network communications to and from host 1202 occur on primary path 1250. Failover mechanism 1210 is configured to monitor primary path 1250 utilizing a heartbeat, polled status register, or other technique. When a failure of primary path 1250 is detected, failover mechanism 1210 may then perform an automatic failover to alternate path 1252. In one embodiment, failover mechanism 1210 may utilize the command "apconfig -P <metaname> -a <interface name>" to perform the failover, where "metaname" is the metaname associated with the logical network connection and "hme1" is the name corresponding to the physical interface. For example, in the system of fig. 12, the command "apconfig -P mether0 -a hme1" may be used to activate alternate path 1252 and inactivate primary path 1250. Alternatively, the "ifconfig" command may be used to configure and bring up the alternate path as described above. Advantageously, high availability and serviceability may be attained in a network with alternate pathing.

Failover with Virtual IP Addresses

Many times an enterprise may wish to have multiple servers, but does not have the budget to invest in additional hardware. A typical example of such a case is an Internet Service Provider (ISP) who supplies multiple Web servers, but does not wish to provide a separate computer system for every server. One well known method of enabling multiple servers on a single machine is through the use of virtual IP addresses. A virtual IP address is a logical address created by a system administrator which typically corresponds to a real interface address. Because the virtual IP address is a logical address, multiple virtual IP addresses may be created which correspond to the same interface. Consequently, multiple Web servers with unique virtual IP addresses may be hosted on a single physical computer. In addition to providing the ability to have multiple servers on a single computer, virtual IP addresses may be used to provide for physical server redundancy and load balancing. For example, if a failure of an interface on a primary server occurs, a failover to a backup server may be utilized to maintain high availability.

Fig. 13 is an example of a computer system utilizing virtual IP addresses. Included in Fig. 13 are hosts 1302A and 1302B. Host 1302A includes failover mechanism 1304A, NIC 1310A, and NIC 1310B. Host 1302B includes failover mechanism 1304B, NIC 1312A, and NIC 1312B. NIC 1310A is coupled to Switch 1320A via interconnect 1330. NIC 1310B is coupled to Switch 1320B via interconnect 1334. NIC 1312A is coupled to Switch 1320A via interconnect 1332. NIC 1312B is coupled to Switch 1320B via interconnect 1336. Switch 1320A is coupled to Switch 1320B via interconnect 1340. NIC 1310A and NIC 1312A are named hme0. NIC 1310B and NIC 1312B are named hme1. NIC 1310A has an interface address of 129.126.1.1 and NIC 1312A has an interface address of 129.126.1.2. In one embodiment, NICs 1310B and 1312B serve as backup interfaces. Backup interfaces NIC 1310B and 1312B may be either cold or warm standby interfaces as described above. As cold interfaces, NIC 1310B and 1312B may have the same MAC address as primary interfaces NIC 1310A and 1312A, respectively.

Failover mechanism 1304A is configured to monitor NIC 1310A and 1310B. Monitoring may be done using a heartbeat mechanism, status register polling, or other similar mechanism. In the event failover mechanism 1304A detects a failure in communications through NIC 1310A, a failover to NIC 1310B may be performed. To

5 perform failover from NIC 1310A to NIC 1312A, failover mechanism transfers the configuration parameters of NIC 1310A to NIC 1310B. Included in the parameters transferred from NIC 1310A to NIC 1310B is the interface address 129.126.1.1. However, because there are a number of virtual addresses associated with NIC 1310A, they must be transferred as well. Failover mechanism 1304 is configured to capture the
10 current virtual IP address configuration of the interface 13310A at the time of failure and configure the backup interface 1310B with the captured parameters. To capture the virtual IP address configuration of an interface, failover mechanism 1304 may utilize a command such as ifconfig or ipconfig. Advantageously, all network IP addresses are failed over and high availability is maintained. In one embodiment, failover mechanism
15 1304 is configured to capture and store the virtual address configurations of the local interfaces upon startup. If upon failure of an interface connection, failover mechanism 1304 is unable to capture the current virtual address configuration, the originally captured startup configuration is used for configuring the backup interface. Similarly, failover mechanism 1304B may be configured to perform a failover from NIC 1312A to NIC
20 1312B.

W In addition to the embodiment described in ^{Fig.} ~~fig.~~ 13, virtual IP addresses may be used in conjunction with a myriad of configurations, including the embodiments described above. When virtual addresses are employed, the included failover mechanism
25 may be configured to failover the virtual addresses as well. Further, while the above description refers to commands “ipconfig” and “ifconfig”, the particular command and syntax used on a particular system may vary. In addition, references to a network “node” or network “device” may include computers, routers, bridges, switches or other similar devices. Finally, the above described embodiments are only intended to be exemplary.

